

PR #3150 完整报告

radixark/miles

test(ci): run the qwen36 session verifier with ROLLBACK, not APPEND_TOOL

合并时间: 2026-09-10 11:02

原文链接: <http://prhub.com.cn/radixark/miles/pull/3150>

执行摘要

- 一句话: 修复 Qwen3.6 端到端测试因工具调用失败模式配置不当导致的间歇性失败。
- 推荐动作: 这个 PR 值得简要阅读, 尤其是其详细的问题分析和验证部分。它展示了如何通过分析模型行为来调整测试配置以解决间歇性失败, 对于处理类似模型与测试框架交互的问题有参考价值。变更本身简单, 但诊断过程深入。

功能与动机

测试 `tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py` 在 scheduled CI 上间歇性失败, 错误为 `AssertionError: Step 6 tool_result exhausted 2 retries after finish_reason='length'`。PR body 中解释, 在 `append_tool` 模式下, 驱动器在工具结果步骤未发现 `tool_calls` 时注入哨兵消息。Qwen3.6 模型在收到此哨兵后既不重试也不作答, 而是在“已调用并回答”和“重试调用”之间反复推理, 直至 `max_tokens`。此行为在约 11-14% 的首次尝试和约 50% 的重试中发生, 导致测试失败。

实现拆解

1. 变更入口: 修改测试配置文件 `tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py`。
2. 核心配置调整: 在 `CONFIG` 字典中, 将 `tool_call_failure_mode` 的值从 `"append_tool"` 改为 `"rollback"`。这是本次变更的唯一逻辑修改点。
3. 变更原因与影响: `rollback` 模式在工具调用失败时会重新采样该轮对话, 而不是注入可能引发模型混淆的哨兵消息。这能有效避免 Qwen3.6 的推理循环, 将步骤完成的 token 消耗从数百至数千大幅降低至几十, 从而消除因 `finish_reason='length'` 导致的间歇性测试失败。此变更仅影响 Qwen3.6 的测试配置, 不影响 Qwen3.5 或其他模型族。
4. 文档 / 注释配套: 在配置行上方添加了详细注释, 解释了切换模式的原因和 Qwen3.6 的具体行为, 便于后续维护。

关键文件:

- `tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py` (模块 Qwen3.6 测试; 类别 test; 类型 test-coverage): 本次变更的唯一文件, 包含了导致测试间歇性失败的配置参数 `tool_call_failure_mode` 及其修复和解释性注释。

关键符号: 未识别

关键源码片段

tests/e2e/sglang/test_session_server_multi_role/test_qwen36.py

本次变更的唯一文件，包含了导致测试间歇性失败的配置参数 `tool_call_failure_mode` 及其修复和解释性注释。

```
CONFIG = ModelConfig(    model_name="Qwen/Qwen3.6-35B-A3B-FP8",
    reasoning_parser="qwen3",    tool_call_parser="qwen3_coder",
    tito_model="qwen36",    tp_size=2,    enable_spec=True,    cycles=2,    # Qwen3.6
    loops on the APPEND_TOOL sentinel: told "the previous turn did    # not emit a
    tool_call, retry" after it already called and answered, it    # re-derives that
    contradiction until max_tokens instead of retrying or    # answering. ROLLBACK
    re-samples the turn instead; a genuine no-tool-call    # failure still surfaces after
    MAX_CONSECUTIVE_TOOL_CALL_FAILURE_ROLLBACKS.
    tool_call_failure_mode="rollback",    # Anthropic tool-call conversion changes raw
    assistant serialization;    # keep this endpoint-only formatting mismatch soft while
    hard gates stay at 0.    anthropic_assistant_text_threshold=1.0,
    anthropic_intermediate_system_expectation="forbidden", )
```

此代码片段展示了修复后的 `CONFIG` 对象定义。关键变更是将 `tool_call_failure_mode` 的值从 `"append_tool"` 改为 `"rollback"`，并在上方添加了多行注释，详细说明了 Qwen3.6 模型在 `APPEND_TOOL` 模式下会陷入循环推理的问题，以及 `ROLLBACK` 模式如何通过重新采样来避免此问题。

评论区精华

审查讨论较少，主要集中在问题的详细诊断和验证上。PR 作者 `guapisolo` 提供了详尽的分析，包括失败日志、对 `append_tool` 和 `rollback` 两种模式行为的对比解释，以及在开发环境中的验证数据（显示切换后 `finish_reason='length'` 事件降为 0，步骤完成 token 大幅减少）。评审者 `Shi-Dong` 直接批准了更改。一个不相关的 Anthropic 端点会话处理异常被发现，但确认与本次变更无关，将单独跟踪。

- Qwen3.6 在 `append_tool` 模式下的非预期行为与测试失败诊断 (correctness): 确认是模型特定行为与测试配置不匹配导致的间歇性失败，而非 `sglang` 或 `Miles` 核心框架的 bug。
- `rollback` 模式作为解决方案的有效性验证 (performance): `rollback` 模式有效解决了由 `append_tool` 哨兵消息引发的 Qwen3.6 循环问题，且测试结果稳定。
- Anthropic 端点出现的无关会话处理异常 (other): 此问题与本 PR 变更无关，将被单独跟踪。

风险与影响

- 风险：低风险。本次变更仅修改一个端到端测试的配置参数，不涉及核心驱动逻辑、调度器或模型代码。主要风险在于配置变更是否能彻底解决间歇性失败。作者提供了充分的验证数据，表明 `rollback` 模式在特定硬件和软件环境下有效。需关注此修复是否在其他 CI 环境（如 AMD MI350）中同样有效。
- 影响：直接影响范围有限：仅影响 Qwen3.6 模型的特定端到端测试 (`test_qwen36.py`) 的运行稳定性。间接影响：提高了 CI 的可靠性，减少了因模型行为导致的假阴性失败。对用户和系统其他模块无直接影响。

- 风险标记：测试配置变更，模型特定行为

关联脉络

- PR #3162 ci: pin the DSv4 bshd/thd parity test to the miles impl: 同为修复 scheduled CI 测试因环境或默认行为变更导致的间歇性失败，属于 CI 稳定性维护系列。
- PR #3148 [CI] temporarily disable failing FT tests: 同为处理 CI 测试失败，本 PR 采用修复配置的方式，而 #3148 采用临时禁用策略，两者都是保持 CI 绿灯的应对措施。